

AN ALTERNATIVE VIEW OF SENSITIVITY IN THE ANALYSIS OF COMPUTER CODES

Michael D. McKay, Richard J. Beckman, Leslie M. Moore and Richard R. Picard
Los Alamos National Laboratory, Los Alamos, New Mexico 87545

KEY WORDS: Sensitivity analysis, uncertainty analysis, risk analysis.

ABSTRACT

This paper addresses the analysis of uncertainty in the output of computer models arising from uncertainty in inputs (parameters). Uncertainty of this type, which is separate and distinct from the randomness of a stochastic model, most often arises when proper input values are imprecisely known. Uncertainty in the output is quantified in its probability distribution, which results from treating the inputs as random variables. The assessment of which inputs are important with respect to uncertainty is done relative to the probability distribution of the output.

1 INTRODUCTION

The evaluation of models in the form of computer codes (computer programs) becomes more important when the models are used in making decisions that have far reaching effects. For example, the complex models used to study global warming, nuclear reactor safety, and environmental safety and restoration provide vital input to regulatory agencies, whose decisions have large impact on our lives. Although models like those used for policy decisions in government vary widely in their mathematical form, they share some important characteristics. Namely, they often “predict” or calculate things one hopes never to observe, for example, serious accidents at nuclear reactors. Secondly, they are functions of many inputs for which costly data collection may be required to determine appropriate values, ranges and so forth. Finally, the relationship between inputs and output is complex.

There are many aspects to the evaluation of the quality of output of a model. The subject addressed in this paper concerns uncertainty in the output attributable to uncertainty in model inputs (or parameters). Within this area, discussion will focus on the sensitivity or importance of the inputs.

2 UNCERTAINTY AND SENSITIVITY

The more traditional, historical approach to sensitivity is founded in the derivative of the output with respect

to each input. As an alternative to numerical calculation, Oblow (1978) and Oblow, Pin and Write (1986) use a technique whereby the capability of calculating derivatives is added to the model using a precompiler called GRESS. Methods from regression are also used, including correlation coefficients, by McKay, Conover and Whiteman (1976), and Iman, Helton and Campbell (1981a, 1981b). Another approach is to look at the output as a random variable and try to find a meaningful decomposition of variance based on the inputs. Fourier methods of Cukier, Levine and Shuler (1978) and, later, Pierce and Cukier (1981) are in this class. In general, these methods appeal to a series expansion of the output, as does the usual propagation of error method which uses a linear expansion of the output in the inputs.

The notion that the variance of the output is a meaningful quantity in assessing importance fits very well with approach taken in this paper, namely, that the importance of inputs can be view with respect to uncertainty in the output. We are interested in the type of uncertainty that can be characterized as being due to the values used for the inputs. A related uncertainty, due to the structure or form of the model itself, is not addressed explicitly. Neither are we concerned with uncertainty due to errors in implementation of the model on a computer. On the other hand, it is certainly acceptable that the calculation might have the randomness of a stochastic process, for which the output of the model is taken as being the cumulative distribution function of the observable output value. In any case, the quantity of interest for uncertainty is the probability distribution of the model output, which is determined by that of the inputs and the transformation of inputs to output via the model. The sensitivity and importance of inputs we want to look at is that relative to the probability distribution of the model output.

3 MATHEMATICAL FRAMEWORK

The uncertainty in the output focused on is that attributable to the inputs. Models often have multiple outputs that can be functions of coordinates of time and location. So as not to needlessly complicate the issue, we consider the case of a single scalar output. Let Y denote the calculated output, which depends on the input vector, X , of length p through the computer model,

$h(\bullet)$. Because proper values of the components of X may be unknown or imprecisely known, or because, in some cases, they can only be described stochastically, it is reasonable to treat X as a random variable and to describe uncertainty about X with a probability distribution. Uncertainty in the calculation Y is captured by its own probability distribution, which is the quantity under study. In summary, then,

$$\begin{aligned} Y &= h(X) \\ X &\sim f_x(x), \quad x \in \mathbb{R}^p \\ Y &\sim f_y(y). \end{aligned} \quad (1)$$

For now, we treat f_x as known, although in practice, knowledge about it is at best incomplete.

We look to the probability distribution, f_y , for answers to the question “What is the uncertainty in Y ?” That is to say, we can use the quantiles of the distribution of Y to construct probability intervals. Alternatively, one might use the variance of Y to quantify uncertainty. In either case, under the assumption that f_y can be adequately estimated, questions answerable with quantiles or moments are covered. However, as has already been mentioned, the issue of how well f_x is known will surely have to be addressed in practice.

Questions of importance of inputs are relative to the probability distribution of Y . That is, they are questions like “Which variables really contribute to (or affect) the probability distribution of the output?” The meaning of importance is given in somewhat of a backwards way as being the complement of unimportant. We say that a subset of inputs is unimportant if the conditional distribution of the output given the subset is essentially independent of the values of the inputs in the subset. These ideas are now examined in more detail.

Suppose that the vector X of inputs is partitioned into X_1 , to be the important components, and X_2 , to be the unimportant ones. Corresponding to the partition, we write

$$\begin{aligned} Y &= h(X) \\ &= h(X_1, X_2). \end{aligned} \quad (2)$$

Furthermore, we assume that X_1 and X_2 are stochastically independent, meaning that

$$\begin{aligned} X_i &\sim f_i(x_i), \quad i = 1, 2 \\ f_x(x) &= f_1(x_1)f_2(x_2). \end{aligned} \quad (3)$$

We address the question of the unimportance of X_2 by looking at the conditional distributions

$$f_{y|x_2} = \text{distribution } Y \text{ given } X_2 = x_2 \quad (4)$$

as compared to f_y , and

$$f_{y|x_1} = \text{distribution } Y \text{ given } X_1 = x_1 \quad (5)$$

for different (all?) values of x_1 and x_2 . We say that X_2 is unimportant if f_y and $f_{y|x_2}$ are not substantially different for all values of X_2 of interest. Similarly, we say that X_1 contains all the important inputs if X_2 is unimportant. Of course, the actual way to compare f_y and $f_{y|x_2}$ must be determined.

Alternatively, comparisons could be made among the distributions $f_{y|x_1}$. Although these distributions are examined, we currently focus on $f_{y|x_2}$ because there is a useful reference distribution, namely, f_y .

The term “screening” is used to mean an initial process of separating inputs X into X_1 , potentially important ones, and X_2 , potentially unimportant ones. The process could resemble a subset selection procedure in regression, in as much as the objective is to select a subset of input variables that “explains” the probability distribution of the output. In the next section, a simple method of partitioning the inputs will be discussed.

4 A SIMPLE SCREENING HEURISTIC

The following is a simple, two-step screening process. The first step is to partition X into a set of “important” components, X_1 , and a set of “unimportant” components, X_2 . The second step is a partial validation to estimate how the components in X_2 actually change $f_{y|x_2}$, to be used to decide if X_2 is really unimportant.

4.1 Partitioning the Input Set

X_2 , a subset of X , is (completely) unimportant when the marginal distribution of Y , equals the conditional distribution of Y given X_2 .

$$f_y = f_{y|x_2} \text{ for all values of } X_2 \quad (6)$$

A way to get an idea of how closely the equality in Eq. (6) holds is through the variance expression Eq. (7) which expresses the marginal variance of f_y in terms of the conditional mean and variance of $f_{y|x_2}$. The variance of Y can be written as

$$V[Y] = E[V[Y | X_2]] + V[E[Y | X_2]]. \quad (7)$$

Equality of the marginal and conditional distributions in Eq. (6) implies that the conditional mean and variance are equal to their marginal counterparts for all values of X_2 . Specifically, the variance (over X_2) of the conditional expectation in Eq. (7) is zero. It is unlikely, of course, that any (realistic set) of the inputs

is completely unimportant. Therefore, the equality between marginal and conditional quantities will be true only in approximation, with the degree of approximation linked to the level of acceptance of the difference between the marginal and conditional distributions of the output, Y .

By inference, if X_1 , the complement to X_2 , is (completely, singly) important, the conditional variance of Y given X_1 is zero, and the variance of the conditional expectation of Y given X_1 is the marginal variance. As before, these relations usually will hold only in approximation. Nevertheless, a comparison of terms in Eq. (7) will offer a way to look at the degree of importance.

The variance decomposition in Eq. (7) suggests a related identity from a one-way analysis of variance, in which the total sum of squares is written as the sum of two components, a “between level” component and a “within level” component. The analysis of variance approach can be used to suggest which components of X belong in X_1 and which in X_2 . We replicate, r times, a Latin hypercube sample (LHS) of size k . The same k values of each component of X will appear in each replicate but the matching within each one will be done independently. The k values correspond to the k levels in the sum of squares decomposition.

In an LHS as introduced by McKay, Conover and Beckman (1979), when the inputs are continuous and stochastically independent, the range of each component of X is divided into k intervals of equal probability content. For a true LHS, a value is selected from each interval according to the conditional distribution of the component on the interval. For this application, it will be sufficient to use the probability midpoint of the interval as the value. The k values for each input are matched (paired) at random to form k input vectors. For the replicates needed in this screening heuristic, r independent combinations of the same values are used to produce the $n = k \times r$ input vectors in total.

A design matrix, M , for an LHS is given in Eq. (8). Each column contains a random permutation of the k values for an input. Each row of the matrix corresponds to a random matching of values for the p inputs used in a computer run.

$$M = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kp} \end{bmatrix} \quad (8)$$

A design matrix for any of the r replicates in this application is obtained by randomly and independently permuting the values in every column of M .

After making the n computer runs using replicated LHS, we begin by looking at the components of X one at a time. Let U denote a component of interest in X , and denote the k values of U by u_1, u_2, \dots, u_k . The n values of the output are labeled by y_{ij} to correspond to the i th value u_i , in the j th replicate (sample). The sum of squares partition corresponding to the input U takes the form

$$\sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y})^2 = r \sum_{i=1}^k (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^r (y_{ij} - \bar{y}_i)^2 \quad (9)$$

$$\text{SST} = \text{SSB} + \text{SSW}$$

where

$$\bar{y}_i = \frac{1}{r} \sum_{j=1}^r y_{ij} \text{ and } \bar{y} = \frac{1}{k} \sum_{i=1}^k \bar{y}_i.$$

A statistic that can be used to assess the importance of U is $R^2 = \text{SSB}/\text{SST}$. Although R^2 is bounded between 0 and 1, the attainment of the bounds is not necessarily a symmetric process. The upper bound is reached if Y depends only on U . In that case, for any fixed value of U , say u_i , the value of Y will also be fixed, making SSW equal to 0. As a result, R^2 will be 1. On the other hand, if Y is completely independent of U , SSB (and, therefore, R^2) is not expected to be 0. We now examine this last point in more detail.

In general, the probability distribution of R^2 will be unknown. To gain insight, however, suppose that we arbitrarily partition a random sample of size n from a normal distribution to form R^2 . (An arbitrary partition would correspond to Y independent of U .) The expected value of R^2 is $(k-1)/(n-1)$, which goes to zero with k/n as n increases. Thus, one might consider $(k-1)/(n-1)$ as a working lower bound associated with a completely unimportant input.

Issues that still need to be addressed include the apportionment of n between r and k , the extension of the design and decomposition to more than one component at a time, and the interpretation of values of R^2 .

Whether or not one uses R^2 or additional methods to develop the sets X_1 and X_2 , there remains the issue of evaluating the partition to see how effective it is in satisfying Eq. (6). In fact, iterating between a partition and validation is what one would do in practice. The next section discusses validation.

4.2 Validation of the Partition

Very simply stated, the validation step looks at X_1 and X_2 and tries to assess how well the partition meets the objective of isolating the important inputs to X_1 . We propose using a very elementary sequence of steps that begins with a sample design resembling Taguchi's (1986) inner array/outer array.

1. Select a sample, S2, of the X_2 s and a sample, S1, of the X_1 s.
2. For each sample element $x_2 \in S2$, obtain the sample of Y corresponding to $\{x_2 \otimes S1\}$.
3. Calculate appropriate statistics for each sample in Step 2, e.g., $\bar{Y}(x_2)$, $s_y^2(x_2)$ and $\hat{F}_{y|x_2}$.
4. Compare the statistics and decide if the difference x_2 makes is acceptable.

The differences seen in the statistics in Step 4 are due only to the different values of x_2 because the sample values for X_1 are the same in each. Hence, the comparisons are reasonable.

The reliability of any validation procedure needs to be evaluated. In this case, S2 may not adequately cover the domain of X_2 , particularly as the dimension of X_2 increases. Merely increasing the size of S2 may not be an acceptable solution if the increase in the number of runs to generate the sample of Y s becomes impossible to accommodate. Inadequate coverage can be due to two reasons. First, regions where the conditional distribution of Y really changes with X_2 alone may be missed. Second, there may be regions where the interaction between X_2 and X_1 in the model has a significant impact on the conditional distribution of Y . Although it has obvious deficiencies, LHS is an appropriate sampling method for generating S2 because it provides marginal stratification for each input in X_2 , meaning that the individual ranges within the components likely have been sampled adequately. Whether or not interaction between X_1 and X_2 will be detected is unknown. As an alternative to LHS, one might use an orthogonal array as described by Owen (1991), which provides marginal stratification for all pairs of input variables.

5 A SIMPLE EXAMPLE

In practice, the model implemented in a computer code will not be expressible in closed form. As an example, however, suppose that the inputs, X , and the output,

Y , are described by

$$\begin{aligned} Y &= b_1 X_1 + b_2 X_2 \\ X_i &\sim N(\mu_i, \sigma_i^2) \end{aligned} \quad (10)$$

Now, suppose that X_2 is presumed unimportant and look at some ramifications of the assumption for this simple model. By unimportant, we mean that the conditional distribution of Y given X_2 is approximately the marginal distribution of Y for all values of X_2 . Because the model in this example is known,

$$\begin{aligned} f_y &= n(b_1 \mu_1 + b_2 \mu_2, b_1^2 \sigma_1^2 + b_2^2 \sigma_2^2) \\ f_{y|x_2} &= n(b_1 \mu_1 + b_2 x_2, b_1^2 \sigma_1^2) \end{aligned} \quad (11)$$

Clearly, the two distributions are not exactly the same for all values of X_2 . Nevertheless, if X_2 is set to its mean value, then the difference in the distributions lies in the variance, and that the marginal variance of Y will be larger than the conditional one. Thus, if X_2 is treated as unimportant and set to its mean value for the purpose of running the code, the effect will be to reduce the variance of the output.

At this stage in partitioning X into X_1 and X_2 , all we know is that $f_{y|x_2}$ is somewhat different from f_y , and that the difference can be restricted to a difference in variances if X_2 is set to its mean. If a suitable cost function can be constructed, one could assess differences between the quantiles of the marginal and conditional distributions.

6 APPLICATION

These methods were applied in a cursory preliminary fashion in the analysis of a compartmental model used to describe the flow of material in an ecosystem. The model calculates concentrations in 15 subsystems, or compartments, as functions of time. For presentation, we have chosen to study the concentration, Y , in one of the compartments at time corresponding to system equilibrium. The flow among compartments, diagrammed in Figure 1, is modeled by a system of linear differential equations. We take as inputs X to the model the 82 constants, called "transfer coefficients," in the equations.

After identifying the model output and inputs, independent beta probability distributions, f_x , were assigned to the inputs. The beta family of distributions was used because of the wide range in shapes it accommodates. We used only unimodal shapes (none of the U-shaped forms) which included symmetric forms and very skewed ones. Parameters of the distributions were inferred from range, best estimate and quantile values obtained from subject-area scientists.

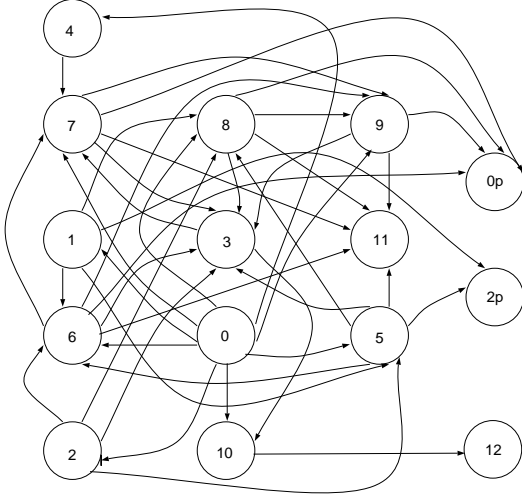


Figure 1. Compartment model

At each stage in the analysis we used a sample size of $n = 1000$ made up of $r = 10$ replicates of an LHS of size $k = 100$, as described in Section 4. We do not suggest either adequacy or minimal sufficiency in the sample size parameters chosen. We picked the numbers hoping that the results would be interpretable, and, if that had not been the case, we would have done something else. We were not concerned with computer time.

The LHS of size $k = 100$ meant that the range of each input was divided into 100 intervals of equal size in probability. Rather than sampling within each interval, we chose to use the interval midpoint as the “sampled” value. The model input vectors for each replicate were constructed by randomly selecting, without replacement, values for each of the 82 inputs. Thus, the replicates differed not in the values used for each individual input, but in the random combinations of 82 values across inputs.

From the first set of 1000 runs, the probability distribution of Y , f_y , created when all 82 inputs are free to vary was estimated. The density function is given in Figure 2 and repeated in Figures 3 and 4 for comparisons. In an iterative manner, we selected inputs as important using R^2 from the sum of squares partition in Eq. (9). In all subsequent iterations, sample values for selected inputs are replaced by nominal values in the input design. The iteration was repeated 4 times, for which 7 of the 82 inputs were selected as possibly being important.

To see how well the selection procedure is working, we look at 2 sets of density functions. First of all, we investigate whether any important inputs have been missed by looking at $f_{y|x_2}$, which describes Y as a function of X_1 (the “important” inputs) for fixed

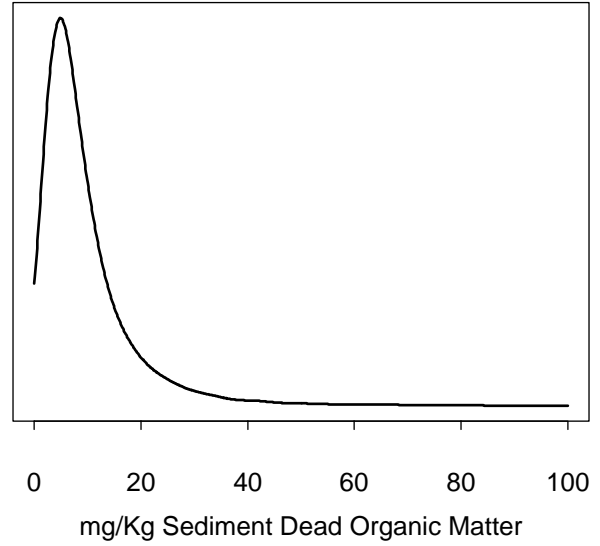


Figure 2. Density function f_y

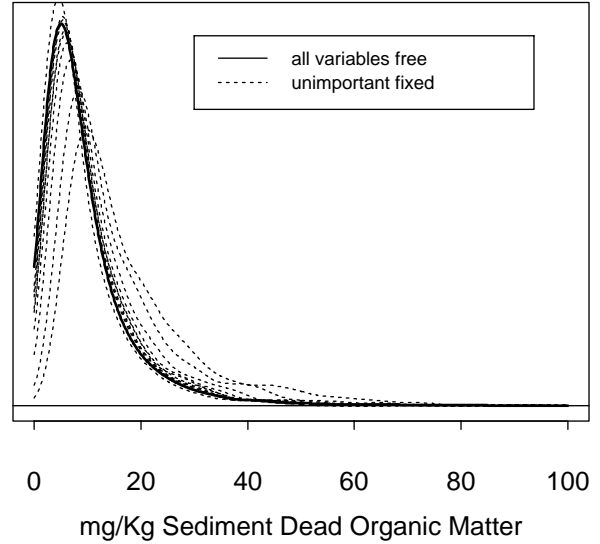


Figure 3. Density functions $f_{y|x_2}$ for 10 values of unimportant inputs X_2

values of X_2 (the “unimportant” inputs). If this density looks like f_y (when all variables are “free”) for all reasonable values of X_2 , we are satisfied no important inputs have been missed. Figure 3 makes the comparison for 10 values of X_2 (actually, from another LHS). The figure indicates acceptable agreement for 8 of 10 values. For 2 of the values of X_2 , the agreement between f_y and $f_{y|x_2}$ is not as close, and further analysis may be prudent.

To see how the set X_1 affects Y , we look at $f_{y|x_1}$ for 10 values of X_1 (from another LHS). These densities are presented in Figure 4. Fixing X_1 produces

densities quite different for the marginal density of Y . We do not know at this point, however, whether the set X_1 contains extraneous inputs.

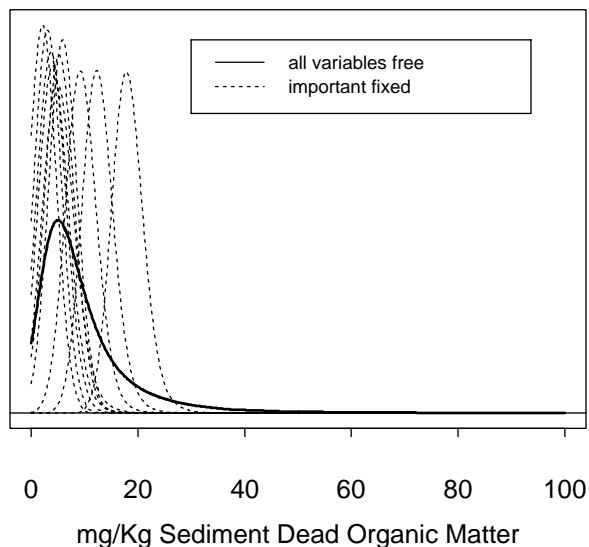


Figure 4. Density functions $f_{y|x_1}$ for 10 values of important inputs X_1

ACKNOWLEDGMENTS

This work was supported by the United States Nuclear Regulatory Commission, Office of Nuclear Regulatory Research. The authors thank A. Juan of the Los Alamos National Laboratory for helpful discussions and ideas.

REFERENCES

- Cukier, R. I., Levine, H. B., and Shuler, K. E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26:1–42.
- Iman, R. L., Helton, J. C., and Campbell, J. E. (1981a). An approach to sensitivity analysis of computer models: Part I—introduction, input variable selection and preliminary variable assessment. *Journal of Quality Technology*, 13(3):174–183.
- Iman, R. L., Helton, J. C., and Campbell, J. E. (1981b). An approach to sensitivity analysis of computer models: Part II—ranking of input variables, response surface validation, distribution effect and technique synopsis. *Journal of Quality Technology*, 13(4):232–240.
- McKay, M. D., Conover, W. J., and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.
- McKay, M. D., Conover, W. J., and Whiteman, D. E. (1976). Report on the application of statistical techniques to the analysis of computer codes. Technical Report LA-NUREG-6526-MS, Los Alamos National Laboratory, Los Alamos, NM.
- Oblow, E. M. (1978). Sensitivity theory for reactor thermal-hydraulics problems. *Nuclear Science and Engineering*, 68:322–337.
- Oblow, E. M., Pin, F. G., and Wright, R. Q. (1986). Sensitivity analysis using computer calculus: A nuclear waste isolation application. *Nuclear Science and Engineering*, 94:46–65.
- Owen, A. B. (1992). Orthogonal arrays for computer integration and visualization. *Statistica Sinica*, 2(2):439–452.
- Pierce, T. H. and Cukier, R. I. (1981). Global nonlinear sensitivity analysis using Walsh functions. *Journal of Computational Physics*, 41:427–443.
- Taguchi, G. (1986). *Introduction to Quality Engineering*. Kraus International Publications, White Plains, NY.